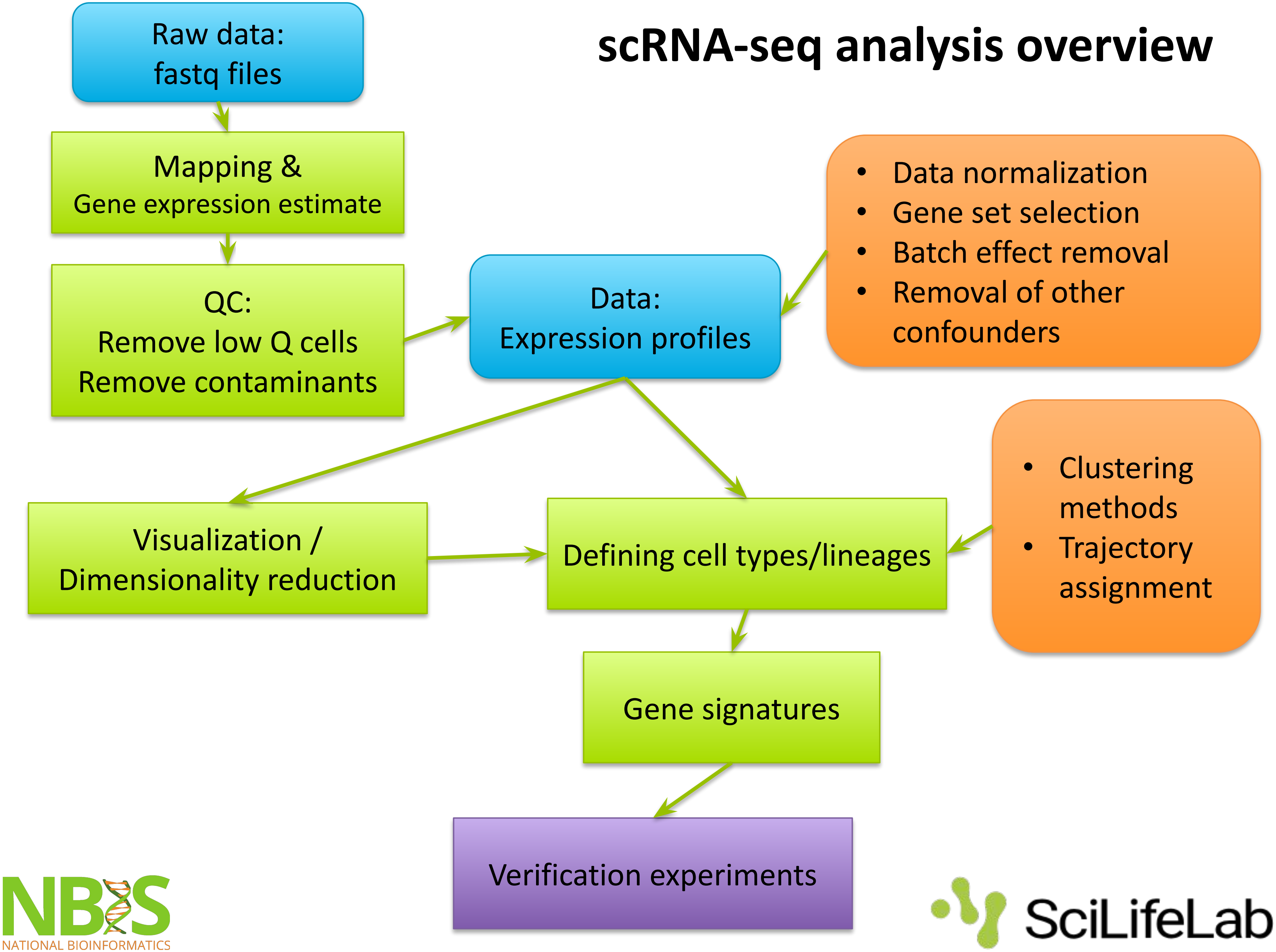
Single cell RNA sequencing data analysis
2026

NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

scRNA-seq analysis overview



Raw data:
fastq files

Mapping &
Gene expression estimate

QC:
Remove low Q cells
Remove contaminants

Data:
Expression profiles

- Data normalization
- Gene set selection
- Batch effect removal
- Removal of other confounders

Visualization /
Dimensionality reduction

Defining cell types/lineages

- Clustering methods
- Trajectory assignment

Gene signatures

Verification experiments

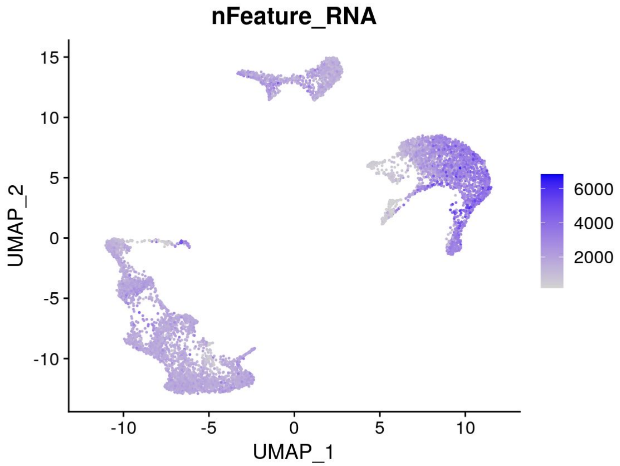
NATIONAL BIOINFORMATICS
INFRASTRUCTURE SWEDEN

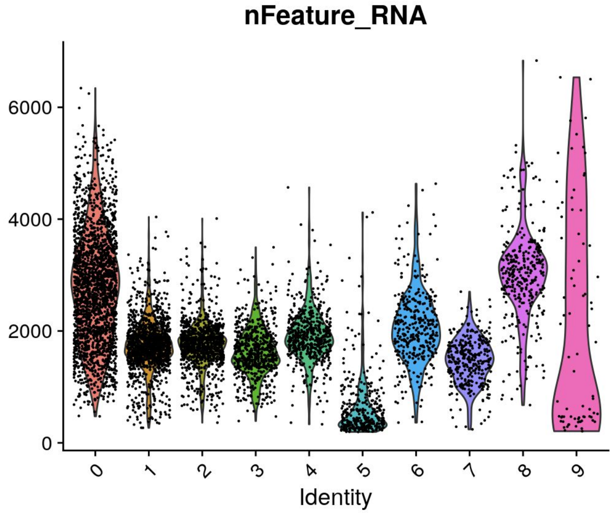
SciLifeLab

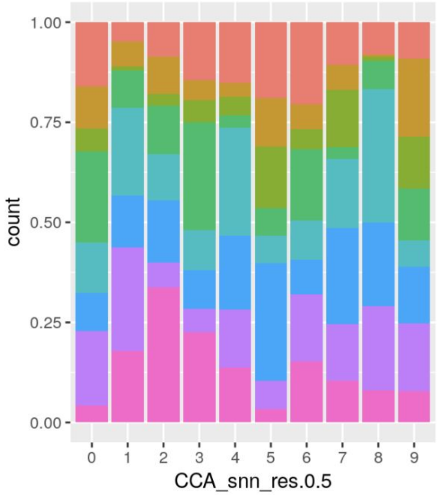
- Data analysis is very seldom a straight line – one pipeline fits all.
  - Often requires several iterations of filtering data, exploring data, refiltering, exploring again, discovering technical artifacts, normalization, exploring again, etc. etc.

- Get to know your data – what types of variation do you have?
  - PCA/UMAP is a good tool for exploring data
- Apply appropriate methods to control for problems that you see.

- Always check for:
  - Batch effects – think of all possible batches.
  - Cell cycle effects if appropriate
  - Separation due to nUMI / nGene / percent mito
- Both at the start of a project and at the end for your final clustering.







- Variable gene selection is a very critical step
  - Filter too much and you may lose populations
  - Keep too much and you may have too much noise

- Similar for choice of PCs

- Clustering – try out a few different approaches
  - Consensus of different methods gives confidence
  - If they do not agree – figure out why!

- Use your biological knowledge to evaluate the results
- Warning! Do not overfit your data to fit your initial hypotheses. Keep an open mind ;-)

- Remember that bioinformatics tools are giving predictions not the truth – always keep a critical mind!
  - Clustering
  - Differential expression
  - Gene set analysis
  - Cell type prediction

- In this course we point out many of the problems that can occur..
- Do not worry too much, in most cases, a standard workflow works well!

- scRNAseq analysis is a fast evolving field with new methods being published all the time.
  - Try to keep up with development
  - **BUT!** You cannot test every new method out there!

Reproducible research in R

- R / Rstudio in Docker containers
  - https://www.andrewheiss.com/blog/2017/04/27/super-basic-practical-guide-to-docker-and-rstudio/
  - https://github.com/rocker-org/rocker
- OBS! On Uppmax/PDC – only Singularity containers are allowed. Most Docker images can be converted.
- Learn more on containers etc:
  - http://nbis-reproducible-research.readthedocs.io/en/latest/
- Rstudio package management – Renv
  - https://rstudio.github.io/renv
- Conda installations of packages – can use conda on both bianca and rackham – `module load conda`

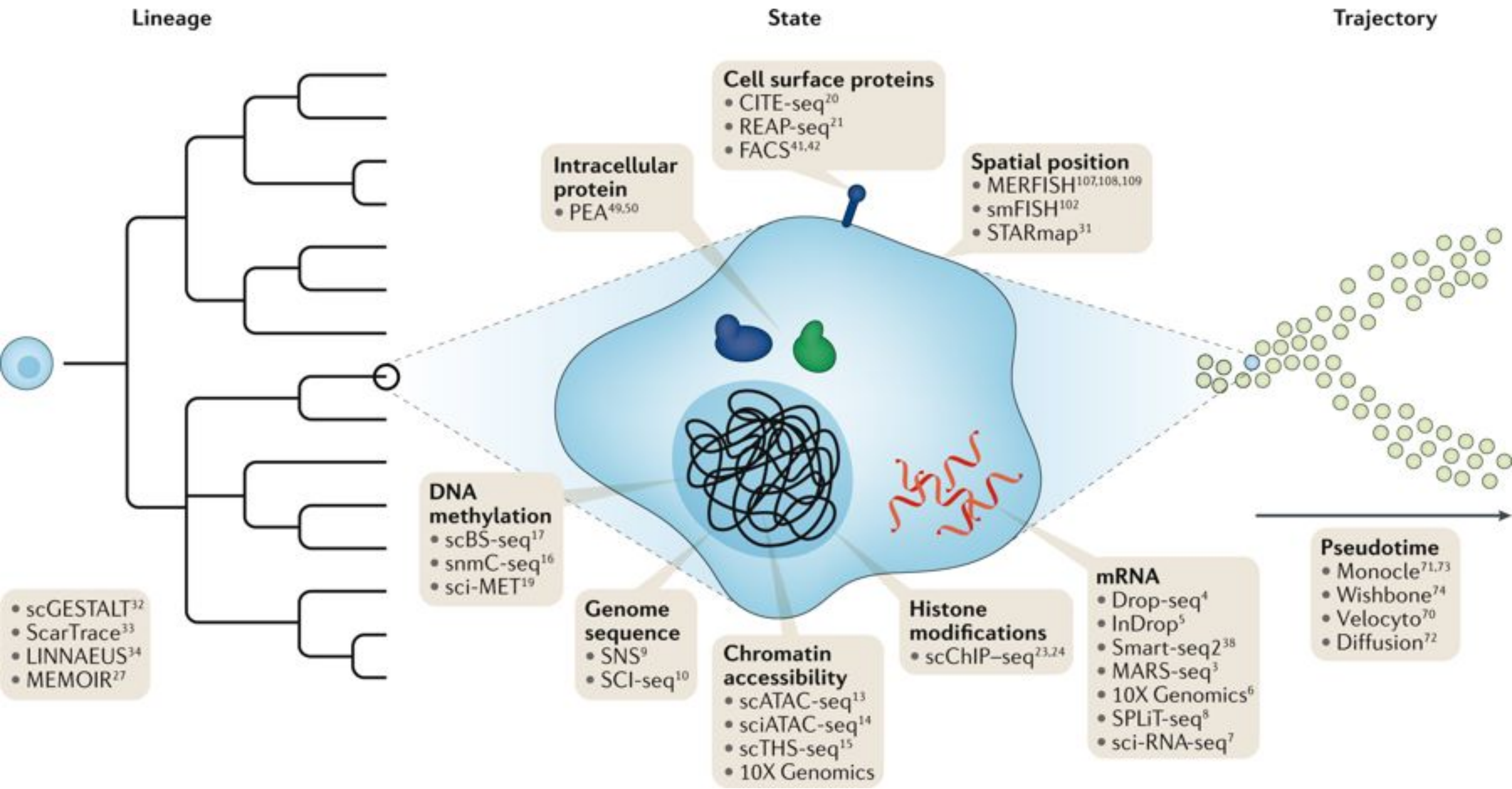
NBIS course in reproducible research:
https://nbisweden.github.io/workshop-reproducible-research/

Compute resources

- In these exercises the datasets were small, but you may have many more cells/samples.
- Structure your code to avoid duplication of matrices and expansion of sparse matrices
  - `rm() & gc()`
- Plan ahead for compute resources, local computer, uppmax or other HPC clusters.
- Human data – raw reads only on encrypted servers like Bianca. Anonymized count matrices are usually fine to analyze in other places.

- We have covered the basic processing, but there is much more you can do…

Single cell omics



Lineage

State

Trajectory

- scGESTALT[32]
- ScarTrace[33]
- LINNAEUS[34]
- MEMOIR[27]

Cell surface proteins
- CITE-seq[20]
- REAP-seq[21]
- FACS[41,42]

Intracellular protein
- PEA[49,50]

Spatial position
- MERFISH[107,108,109]
- smFISH[102]
- STARmap[31]

DNA methylation
- scBS-seq[17]
- snmC-seq[16]
- sci-MET[19]

Genome sequence
- SNS[9]
- SCI-seq[10]

Chromatin accessibility
- scATAC-seq[13]
- sciATAC-seq[14]
- scTHS-seq[15]
- 10X Genomics

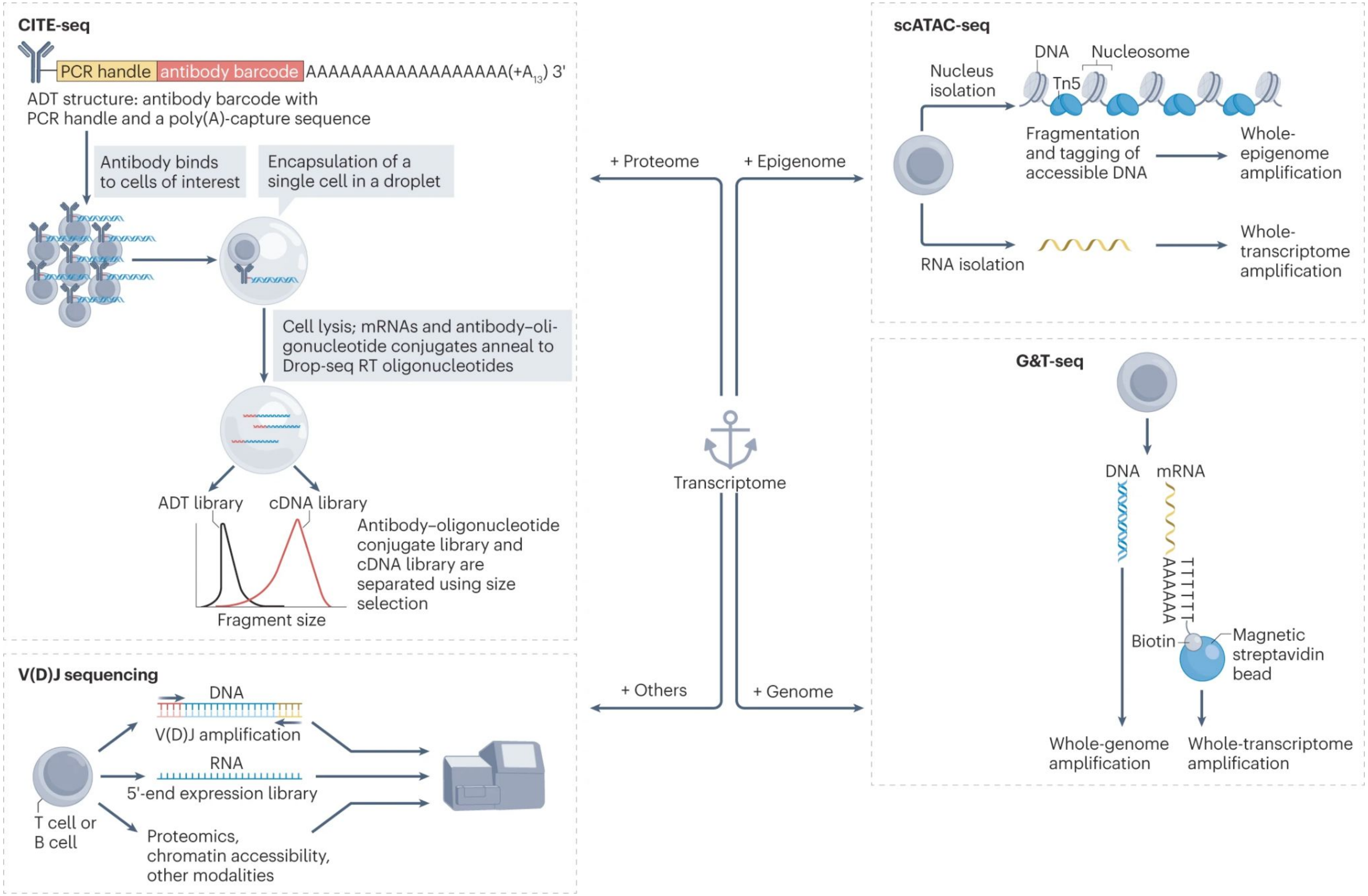
Histone modifications
- scChIP–seq[23,24]

mRNA
- Drop-seq[4]
- InDrop[5]
- Smart-seq2[38]
- MARS-seq[3]
- 10X Genomics[6]
- SPLiT-seq[8]
- sci-RNA-seq[7]

Pseudotime
- Monocle[71,73]
- Wishbone[74]
- Velocyto[70]
- Diffusion[72]

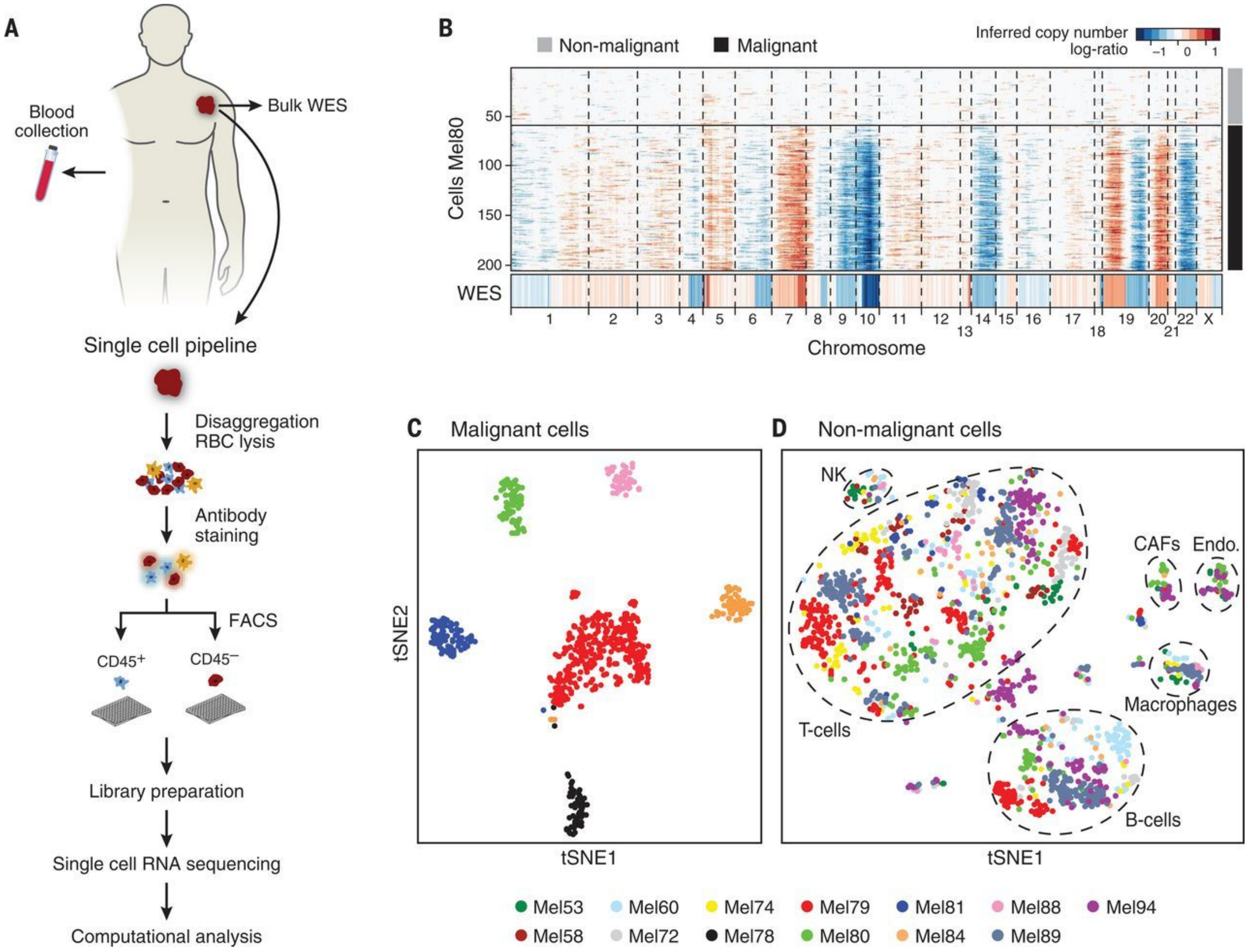
(Stuart & Satija, Nature Rev. Genetics 2019)

SC Multimodal omics



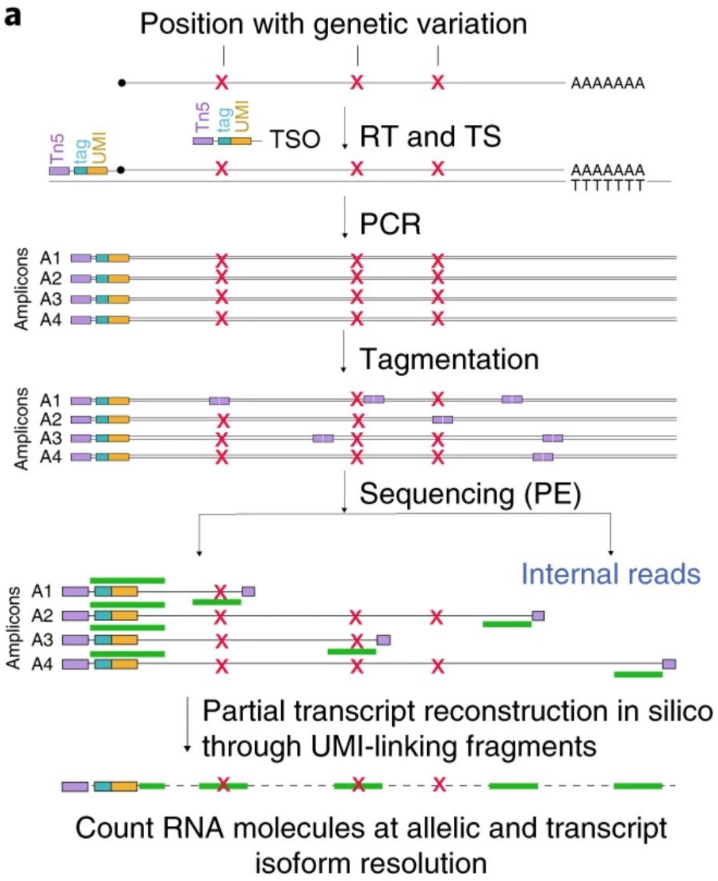
(Baysoy et al, Nat Rev Mol Cell Biol, 2023)

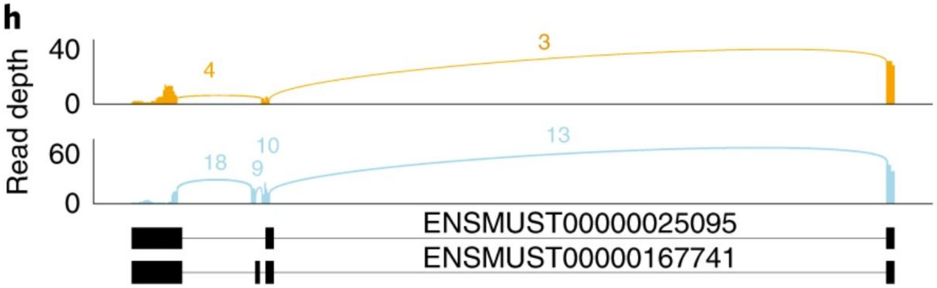
Copy-number variation (CNV) profiling with RNAseq



(Tirosh et al. *Science* 2016)

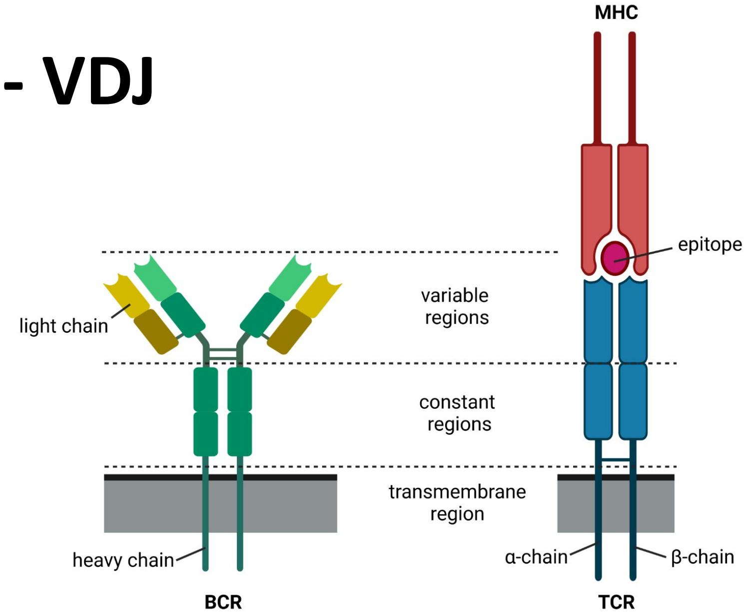
Allele and isoform information with SmartSeq3

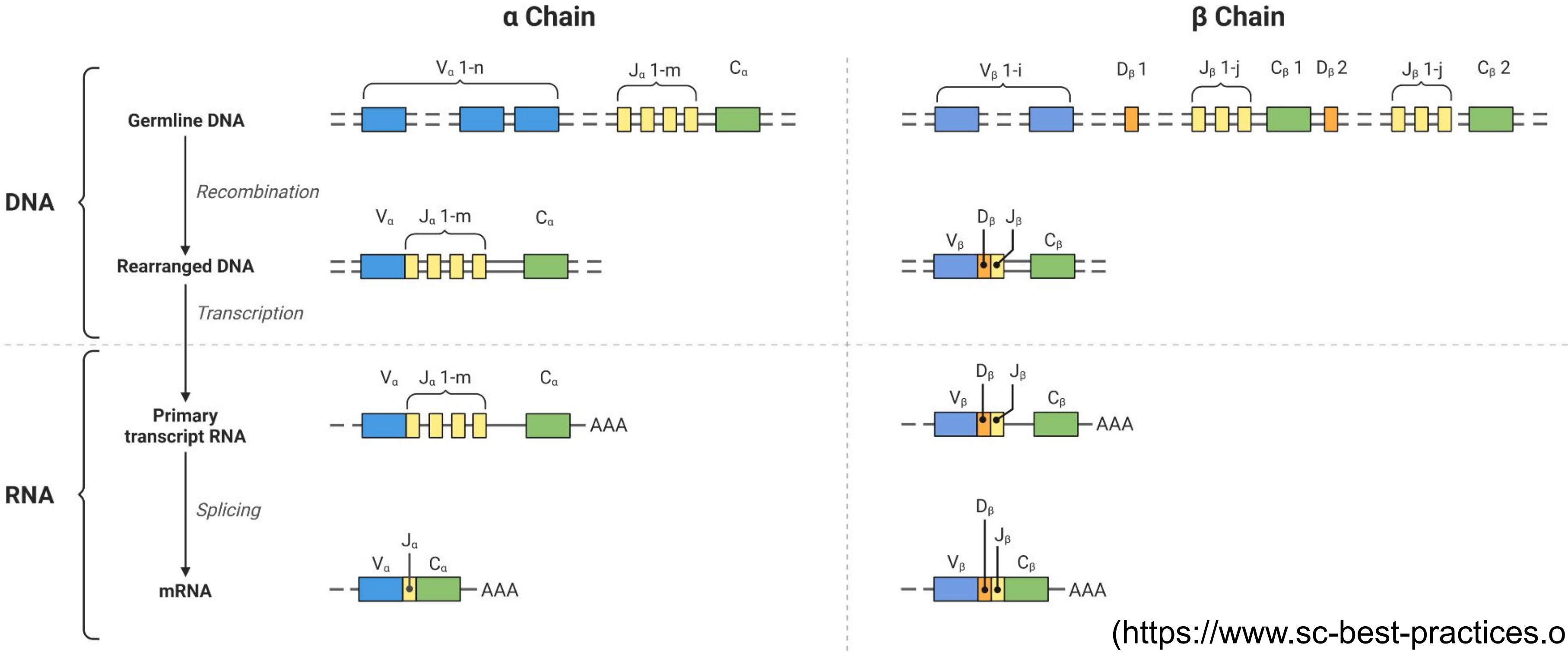




(Hagemann-Jensen et al. Nat. Biotech 2020)

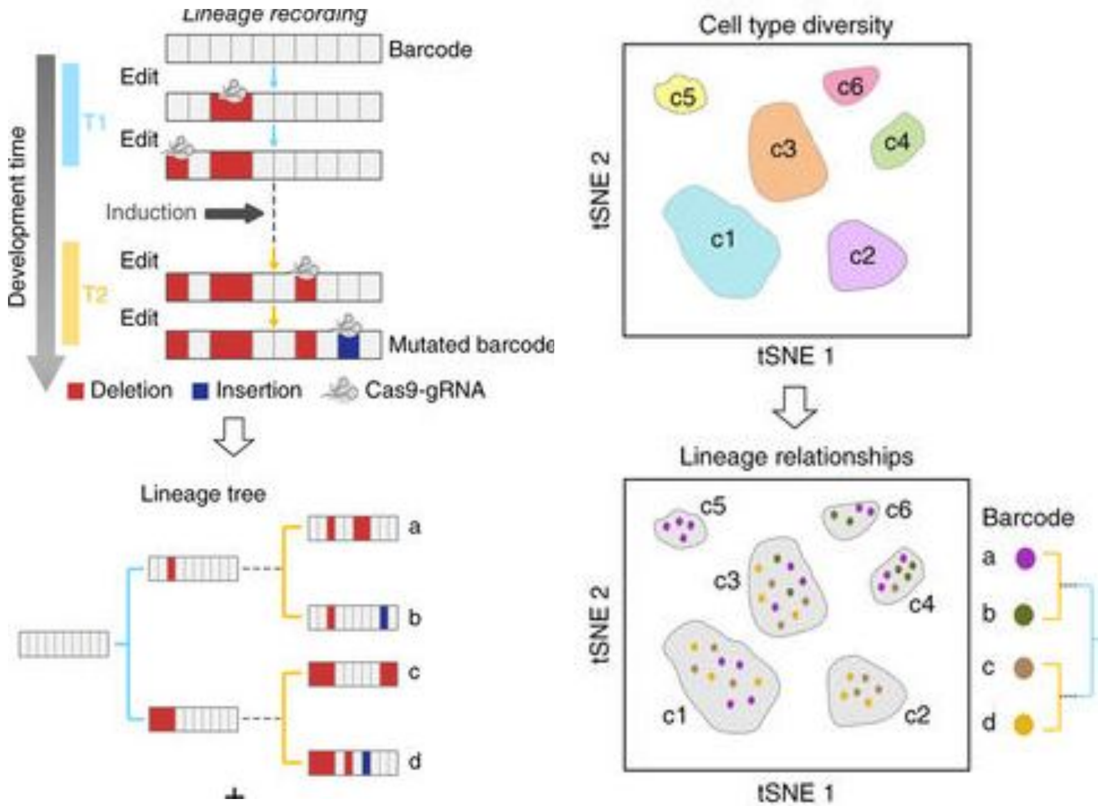
Immune receptor repertoire - VDJ

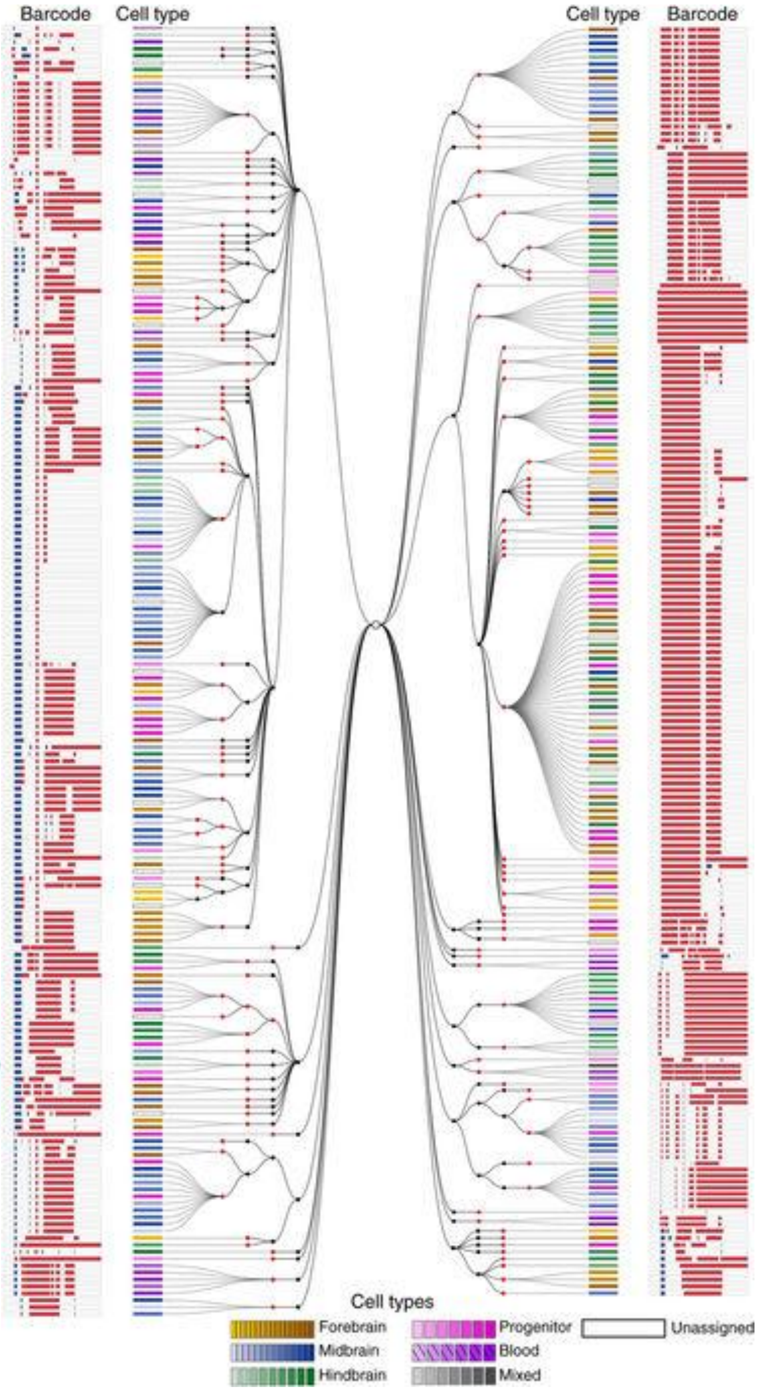




(https://www.sc-best-practices.org)

scGESTALT – lineage tracing and cell profiling with CRISPR-Cas9 editing of barcodes



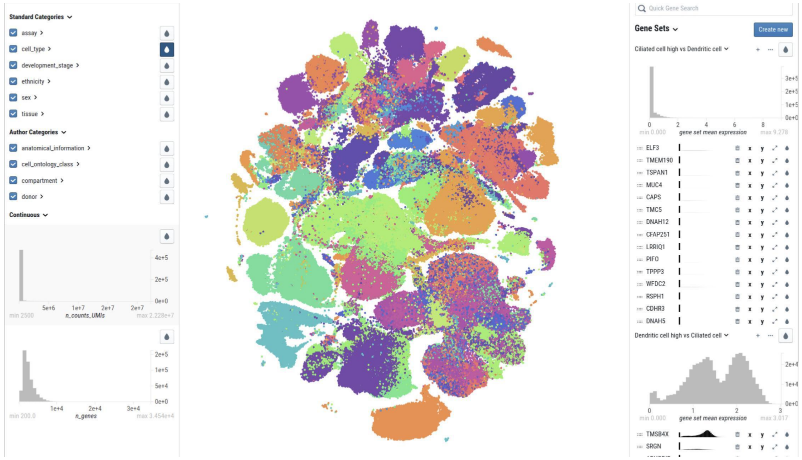


(Raj et al. Nature Biotech 2018)

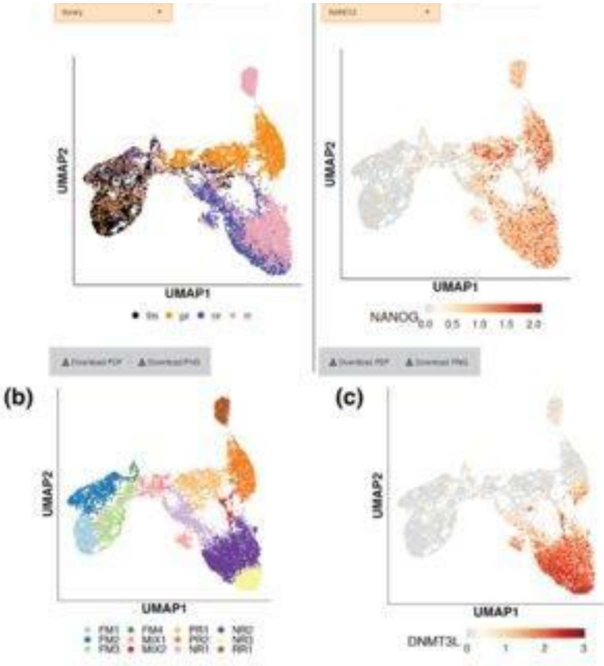
NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Interactive visualization

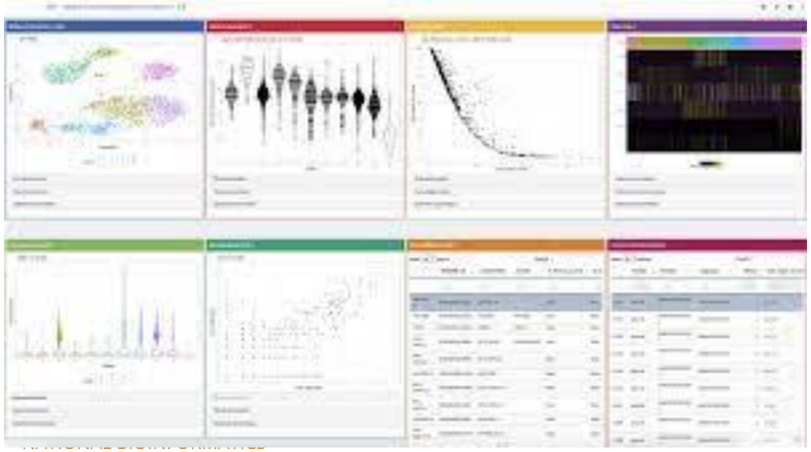
Cellxgene



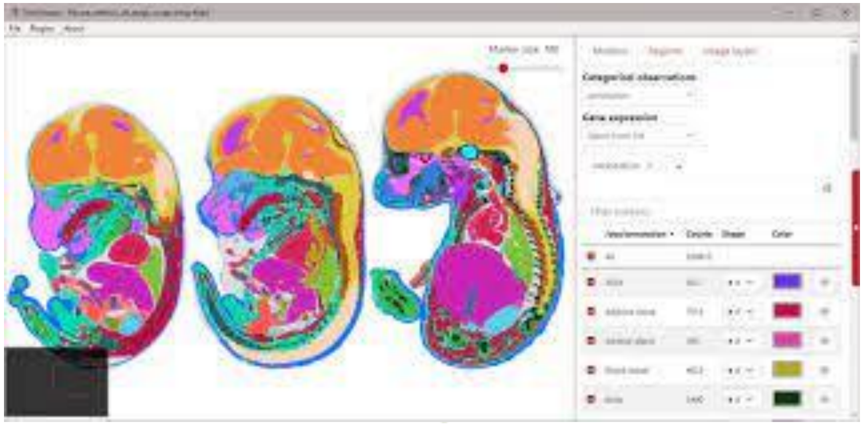
Shinycell



iSEE



TissUUmaps



Some resources

- Course at: https://hemberg-lab.github.io/scRNA.seq.course/
- Scanpy course: https://www.sc-best-practices.org/
- Orchestrating Single-Cell Analysis with Bioconductor http://bioconductor.org/books/3.13/OSCA/
- Many of the packages have good tutorials on their websites
- Repo with scRNA-seq tools: https://github.com/seandavi/awesome-single-cell

Need help?

- NBIS project support
- Courses in programming and other types of analyses.
- Drop-in sessions every Tuesday 14.00
- More info at: http://nbis.se/

NATIONAL BIOINFORMATICS
INFRASTRUCTURE SWEDEN

Please fill in the Evaluation Form

Your feedback is important so that we can help improve the course.

NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Bring Your Own Data

BYOD April 27th

9:00 - Short introduction

9:15 - Data managers present

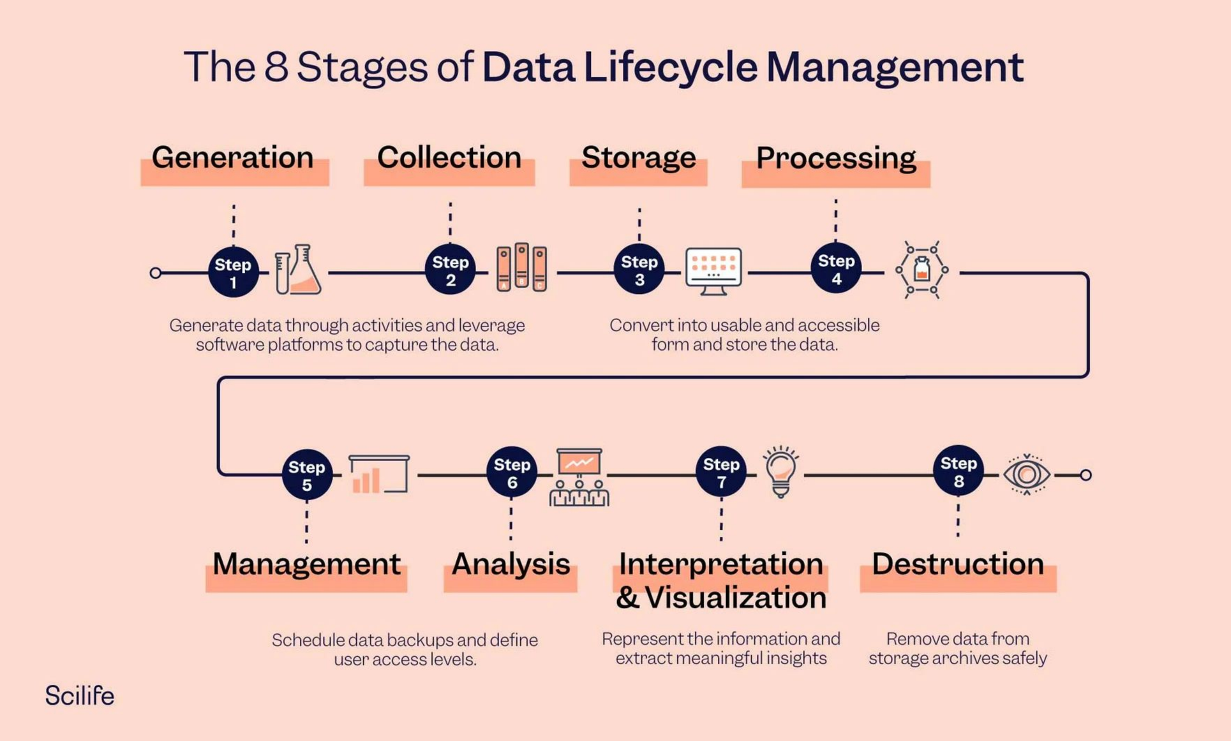
You work on your own and we will be around to help.

14:30 - Presentations of all your projects.

16:00 - Summary of the day.

NATIONAL BIOINFORMATICS
INFRASTRUCTURE SWEDEN

Data Management



The 8 Stages of **Data Lifecycle Management**

Generation — Step 1

Collection — Step 2

Generate data through activities and leverage software platforms to capture the data.

Storage — Step 3

Processing — Step 4

Convert into usable and accessible form and store the data.

Management — Step 5

Analysis — Step 6

Schedule data backups and define user access levels.

Interpretation & Visualization — Step 7

Represent the information and extract meaningful insights

Destruction — Step 8

Remove data from storage archives safely

Scilife

NBIS Data Managers - our guides through the data life cycle will have a short presentation and be present for consultations during the day.

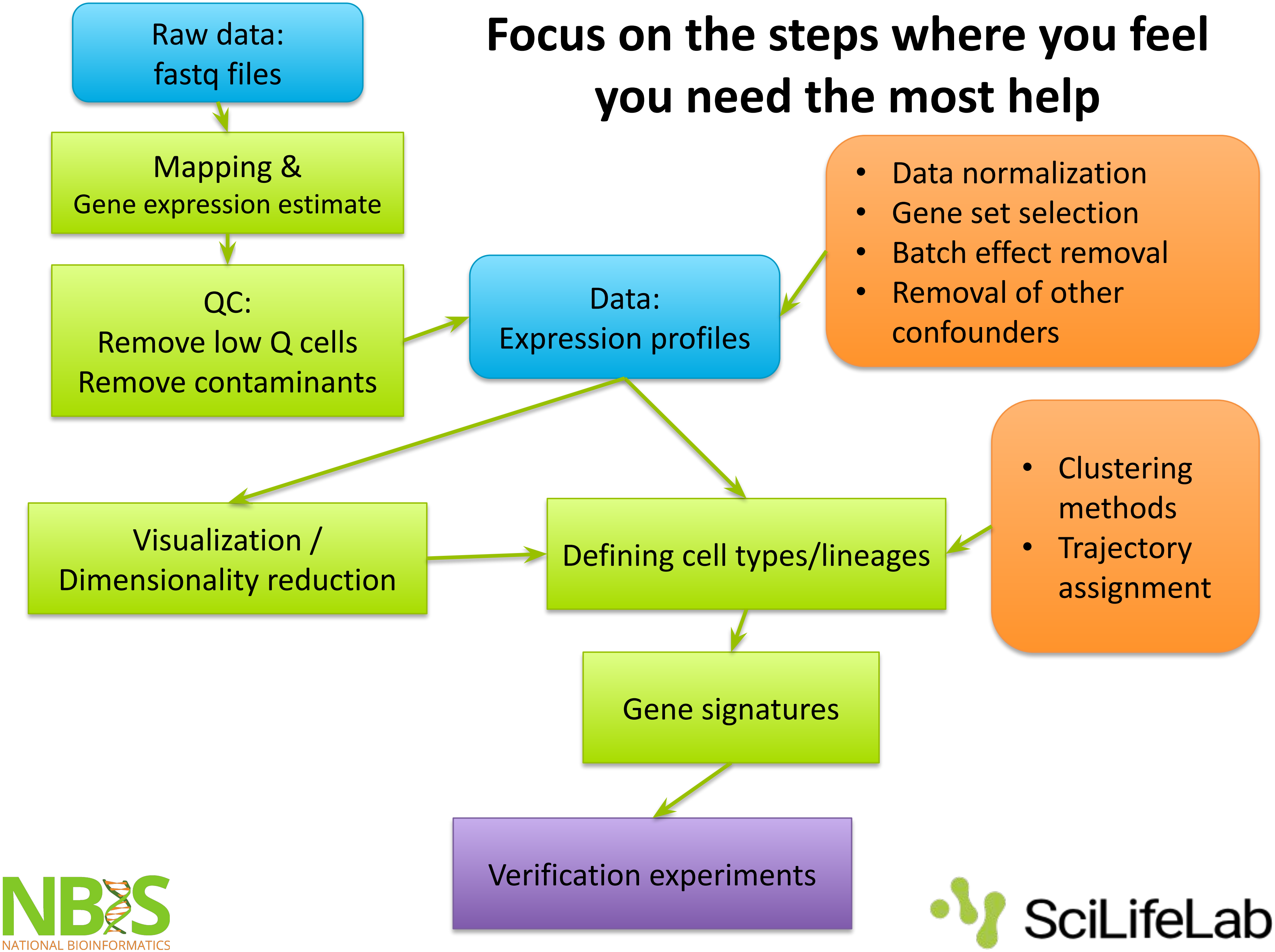
NBIS
NATIONAL BIOINFORMATICS
INFRASTRUCTURE SWEDEN

SciLifeLab

Datasets

- If you have your own data - use that!
- If not - select a dataset from a tissue/celltype that you are interested in.
- Do not select a very large dataset, will be too much computational time during one day!
  - Subsample the cells or select another dataset.
- **If you have sensitive human data, make sure you have permission to work on count matrices locally or on the server.**

Focus on the steps where you feel you need the most help



Raw data:
fastq files

Mapping &
Gene expression estimate

QC:
Remove low Q cells
Remove contaminants

Data:
Expression profiles

- Data normalization
- Gene set selection
- Batch effect removal
- Removal of other confounders

Visualization /
Dimensionality reduction

Defining cell types/lineages

- Clustering methods
- Trajectory assignment

Gene signatures

Verification experiments

NBIS
NATIONAL BIOINFORMATICS
INFRASTRUCTURE SWEDEN

SciLifeLab

Breakout rooms

Please answer our short survey (link also in slack)

https://docs.google.com/forms/d/e/1FAIpQLSeAr90_n7VmgxhZDq5fIhY_XW6CbSoFcrOddNMaRh0UEu3eWw/viewform

We will try to make groups that make sense.

If you know that you won’t attend: Please fill in the survey to let us know!

Compute environments

- If you have a small dataset - less than 50K cells, probably okay to work on the Serve system. Upload it using the Serve File Manager
- For large datasets, use one of the following options:
  - Downsample the dataset and upload to Serve
  - Use the course containers and Docker locally
  - Use a local setup (at your own risk)
    - Conda-files are available on GitHub but not tested outside of the containers
    - Packages can be installed as you go, but different versions can cause issues with dependencies
    - If you haven't already - consider reproducibility! https://uppsala.instructure.com/courses/107067
- Make sure it runs before April 27th
- Ask in the slack channel **#precourse** if you have problems.

NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

More information

https://nbisweden.github.io/workshop-scRNAseq/other/containers.html

- Option A - running labs on Serve.
  - Use the File Manager at the bottom of the project dashboard to upload your data.
- Option B - running labs locally.

Presentation

We want a **very** short - max 5 minutes presentation of what you have done.

- 1 slide
- What dataset
- Which analysis steps did you perform?
- Which were your main results?
- What were the biggest problems you encountered?

NATIONAL BIOINFORMATICS
INFRASTRUCTURE SWEDEN

QUESTIONS?

NBIS
NATIONAL BIOINFORMATICS
INFRASTRUCTURE SWEDEN

SciLifeLab